MULTIPLE CHOICE QUESTION ANSWERING BY FINE TUNING BERT

Group 2: Manikanta Mandlem, Meetu Patel, Nihaal Subhash, Sai Indraneel Amara

The Dataset

OpenBookQA is a new kind of question-answering dataset modeled after open book exams for assessing human understanding of a subject.

It consists of 5,957 multiple-choice elementary-level science questions (4,957 train, 500 dev, 500 test), which probe the understanding of a small “book” of 1,326 core science facts and the application of these facts to novel situations.

For training, the dataset includes a mapping from each question to the core science fact it was designed to probe. Answering OpenBookQA questions requires additional broad common knowledge, not contained in the book.

The Task

Train two models:

1. Fact = 0. This model will have no associated fact with it. It must predict the correct answer amongst the 4 options with common sense and its self taught knowledge.
2. Fact = 1. This model will have an associated relevant fact with the question. It can use this relevant fact as well while predicting the answer.

Approaches

1. Using some explicit method to eliminate one of the answer choice and retraining your model to perform multiple-choice question-answering with three choices.
2. Using external knowledge from sources to fine tune then incrementally fine tune on OpenBookQA.

Approach 1: Eliminating One Option

1. We attempted three different approaches to eliminate one option
    a. Eliminate the option with the highest perplexity of GPT2
    b. Use the dedicated lmfit library to eliminate one option using GPT2 by calculating the product of the probabilities of a sentence, followed by normalizing for sentence length
    c. Use the dedicated lmfit library to eliminate one option using GPT2 Large by calculating the product of the probabilities of a sentence, followed by normalizing for sen

Approach 1: Results

By default, randomly eliminating one option out of 4 given options, you have a probability of 75% to NOT eliminate the correction option. The proposed approaches gave us the following results:

Sr. No.	Approach	Accuracy (Probability of the removed option not being the correct option)
1	Random (Baseline)	75%
2	Remove using perplexity	73%
3	Remove using GPT2 product of token probabilities normalized by sentence length	82%
4	Same as (3) but with GPT2 Large	80%

Approach 1: Results

We selected the ‘Remove using GPT2 product of token probabilities normalized by sentence length’ approach as it had the highest probability (82%) of not removing the correction option. After that we trained our current model to choose the correct option amongst the three options. The results are as follows:

Sr. No.	Approach	Fact = 0	Fact = 1
1	Baseline	0.550	0.691
2	Explicitly removed 1 option	0.453	0.518

Approach 2: Fine tune BERT using additional data and then incrementally fine tune on additional data:

1. We attempted to fine tune using 3 different datasets:
    a. SWAG
    b. OpenBookQA
    c. Allen AI Science Challenge

SWAG

Given a partial description like "she opened the hood of the car," humans can reason about the situation and anticipate what might come next ("then, she examined the engine").

SWAG (Situations With Adversarial Generations) is a large-scale dataset for this task of grounded commonsense inference, unifying natural language inference and physically grounded reasoning.

The dataset consists of 113k multiple choice questions about grounded situations. Each question is a video caption from LSMDC or ActivityNet Captions, with four answer choices about what might happen next in the scene, with only one being correct.

CommonsenseQA

CommonsenseQA is a multiple-choice question answering dataset that requires different types of commonsense knowledge to predict the correct answers.

It contains 12,102 questions with one correct answer and four distractor answers.

We modified this dataset and removed one distractor since we wanted the model to work on picking one option out of 4 rather than one option out of 5.

We also fine tuned the model on train, test and valid of CommonsenseQA because we wanted it to learn as much of common sense knowledge as possible. After that we incrementally fine tuned on the OpenBookQA dataset.

Allen AI Science Challenge

A dataset of 7,787 genuine grade-school level, multiple-choice science questions, assembled to encourage research in advanced question-answering.

The dataset is partitioned into a Challenge Set and an Easy Set, where the former contains only questions answered incorrectly by both a retrieval-based algorithm and a word co-occurrence algorithm.

The ideology behind selecting this dataset was that it would help imbibe some general knowledge to the model which would assist the model in common sense reasoning tasks, especially when the relevant fact is not provided.

Approach 2: Results

Sr. No.	Approach	Fact = 0	Fact = 1
1	Baseline	0.550	0.691
2	Fine Tuning on Pooled Swag and OpenBookQA	0.560	0.692
3	Fine Tuning on Swag and incrementally fine tuning on OpenBookQA	0.558	0.691
4	Fine Tuning on CommonsenseQA and incrementally fine tuning on OpenBookQA	**0.584 (Best Result)**	**0.710 (Best Result)**
5	Fine Tuning on Allen AI Science Challenge (Easy) and incrementally fine tuning on OpenBookQA	0.550	0.708
6	Fine Tuning on Pooled CommonsenseQA + Allen AI and incrementally fine tuning on OpenBook QA	0.564	0.700

Future work

- Use the representations generated by GPT-2 (eg: the context embeddings or logits) to learn a classifier to eliminate one incorrect choice.
- Train a model to look at the question and retrieve the relevant fact from the list of facts in crowdsourced-facts.txt

THANK YOU

Manikanta Mandlem
Meetu Patel
Nihaal Subhash
Sai Indraneel Amara